"Statistically interpretable importance indices for Random Forests"

Paul, Jérôme ; Dupont, Pierre

Abstract

In the original Random Forest (RF) approach, Breiman proposes an embedded feature importance index. It is proportional to the decrease in tree accuracy, estimated on the out-of-bag (OOB) samples, when permuting a particular variable. Such multivariate index takes into account the interactions between variables but is not straightforward to interpret in a statistical sense. In particular, it is hard to decide which variables are statistically significant and, specifically, to assign a p-value to such a decision. We proposed a statistical procedure to measure variable importance, that tests if variables are significantly useful in combination with others in a forest in (Paul et al. 2013). The importance J_χ^2 of a variable is defined as the p-value, corrected for multiple testing, that the tree class vote distribution changes when permuting the feature. Those changes are estimated on the OOB samples and assessed by a Pearson's chi squared test. The outputted p-values offer a natural t...

Document type : *Communication à un colloque (Conference Paper)*

Statistically interpretable importance indices for Random Forests

Jérôme Paul JEROME.PAUL@UCLOUVAIN.BE
Pierre Dupont PIERRE.DUPONT@UCLOUVAIN.BE
Université catholique de Louvain – ICTEAM/Machine Learning Group,
Place Sainte Barbe 2 bte L5.02.01, B-1348 Louvain-la-Neuve, Belgium

Keywords: Feature selection, variable importance index, Random Forests, statistical test

In the original Random Forest (RF) approach, Breiman (2001) proposes an embedded feature importance index. It is proportional to the decrease in tree accuracy, estimated on the out-of-bag (OOB) samples, when permuting a particular variable. Such multivariate index takes into account the interactions between variables but is not straightforward to interpret in a statistical sense. In particular, it is hard to decide which variables are statistically significant and, specifically, to assign a p-value to such a decision.

We proposed a statistical procedure to measure variable importance, that tests if variables are significantly useful in combination with others in a forest in (Paul et al., 2013). The importance J_{χ^2} of a variable is defined as the p-value, corrected for multiple testing, that the tree class vote distribution changes when permuting the feature. Those changes are estimated on the OOB samples and assessed by a Pearson's χ^2 test. The outputted p-values offer a natural threshold to decide which features are statistically significant in combination with the other features in the forest. Experiments conducted on synthetic and real high-dimensional datasets show that J_{χ^2} correctly identifies relevant variables provided a large number of trees (typically $10,000$). The feature ranking is also largely correlated with Breiman's index.

Further analyses (Paul & Dupont, 2014) compare J_{χ^2} to two alternative procedures proposed in (Huynh-Thu et al., 2012): 1Probe and mr-Test. They also allow to convert Breiman's importance into p-values. However, they are conceptually and computationally more complex than J_{χ^2}. Indeed, they resort on a resampling process that repeatedly builds RFs in order to estimate a null importance distribution from which p-values can be computed. Since J_{χ^2} is directly estimated on the OOB samples with no need of additional resamplings, it appears that it requires an order of magnitude less trees than the two other approaches to yield similar sets of selected variables.

One potential drawback of J_{χ^2} is that the assumption of independence of the χ^2 test may become strongly violated when growing forest with exceedingly many trees while the number of independent samples in the dataset is left unchanged. One can address this potential issue by considering a Kolmogorov-Smirnov test while providing similar results to J_{χ^2}. We also evaluate here alternative statistical procedures based on the tree accuracy distributions with and without permuting variables.

To sum up, we study several RF feature importance indices with the objective of relating them to well defined statistical tests. Such relation offers a statistical interpretation of those indices after translating them into $p-$values. It also provides a natural threshold to highlight relevant variables. Practical experiments, both on artificial and real data from DNA microarrays, show that one is able to retrieve important variables while drastically reducing the computational cost of recently proposed alternatives.

References

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Huynh-Thu, V. A. A., Saeys, Y., Wehenkel, L., & Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics (Oxford, England)*, *28*, 1766–1774.

Paul, J., & Dupont, P. (2014). Inferring statistically significant features from random forests. *Under review*.

Paul, J., Verleysen, M., & Dupont, P. (2013). Identification of statistically relevant features from random forests. *ECML-PKKD 2013 workshop proceedings, Solving Complex Machine Learning Problems with Ensemble Methods* (pp. 69–80). Prague, Czech Republic.